AWS ML Associate — Section 2.2: Feature Engineering and Selection

This section covers creating, transforming, and selecting features to improve model accuracy and efficiency. It focuses on encoding, scaling, feature extraction, dimensionality reduction, and feature store usage.

# Feature Engineering

• Encoding categorical variables — one-hot encoding, embeddings

• Scaling numerical features — min-max, z-score normalization

• Text features — Bag-of-Words, TF-IDF, embeddings

• Image features — pixel normalization, transfer learning (CNNs)

# Feature Selection

• Filter methods — correlation, chi-square test

• Wrapper methods — recursive feature elimination (RFE)

• Embedded methods — regularization (L1/Lasso), tree-based feature importance

• Dimensionality reduction — PCA, t-SNE (mainly visualization)

# AWS Tools

• SageMaker Data Wrangler — feature transformations

• SageMaker Processing — custom preprocessing scripts

• SageMaker Feature Store — central repository for consistent features

# Exam Tips

• Scaling is critical for distance-based algorithms (k-means, k-NN, SVM)

• One-hot encoding is standard for categorical data

• PCA reduces dimensionality, but interpretability may be lost

• Use Feature Store for reusability across train/inference